Mini Project #5

Pranil Kamble, Gaurav Madnani

Both worked together to finish the two questions. We both worked to find the solution of the questions asked. Both worked to make the code efficient and gave annotations to the code. Pranil worked on report and reasoning asked for question 1. Gaurav worked on report and reasoning for the questions asked for question 2. Both partners worked efficiently to complete the project requirements.

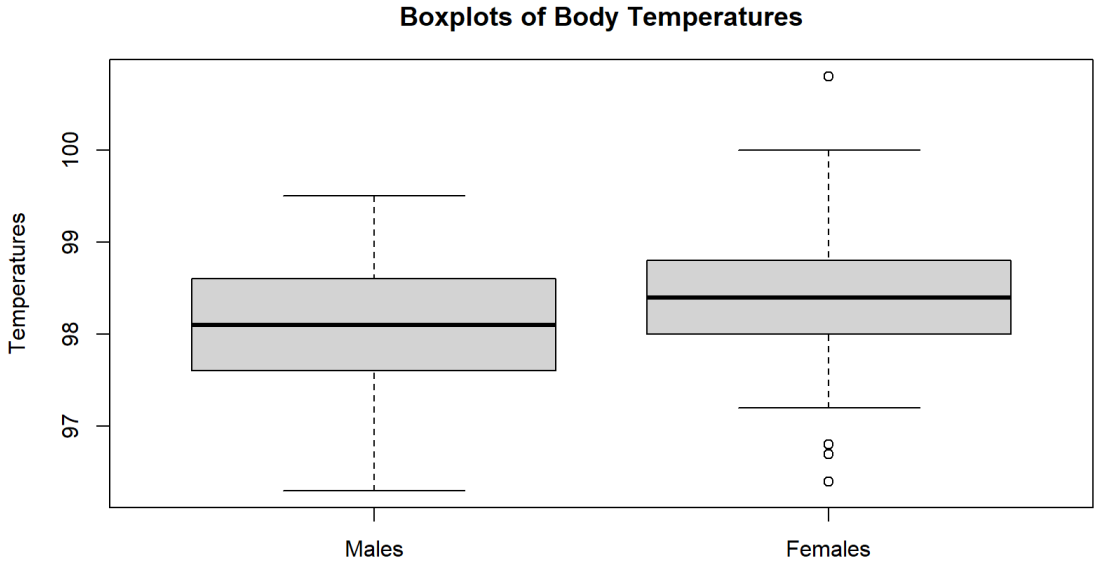
Section 1

Question 1

The dataset contains 3 columns body_temperature, gender, and heart_rate. Load dataset using read.csv function then separate male and female data for analysis.

a) Do males and females differ in mean body temperature? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

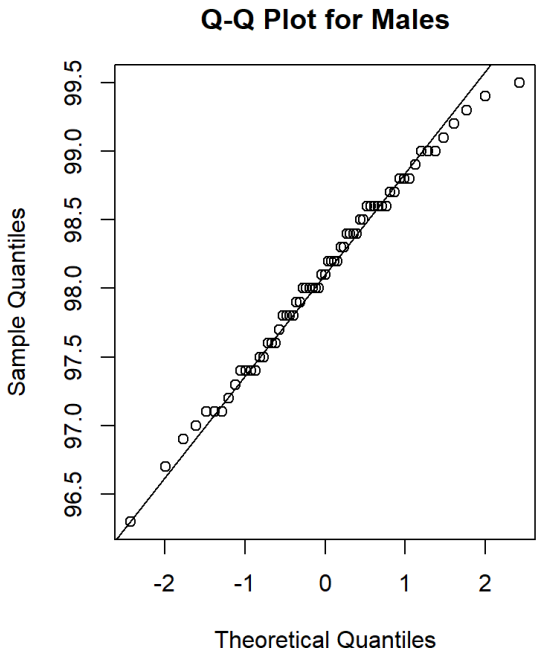
To compare mean body temperature of male and female we can plot a box plot.

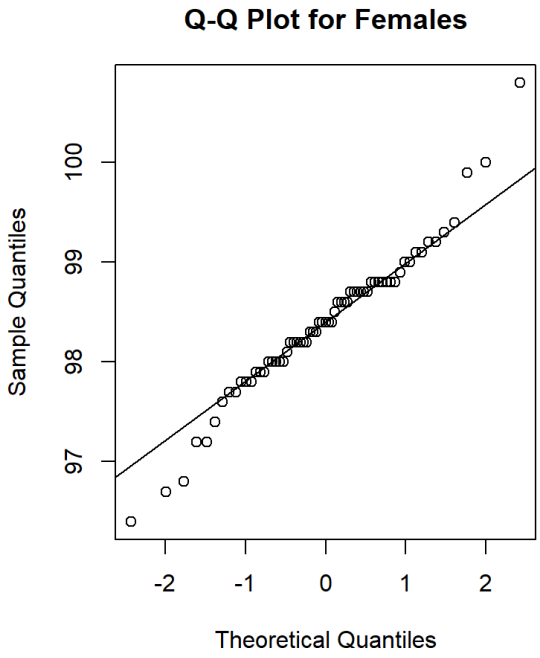


From the boxplot we can observe that body temperature of female has slightly higher quartile values and mean value. There are more outliers in female data which implies there is more variability.

Let's plot a QQ plot to see the normal distribution of both male and female body temperature.

We can observe that the distribution of body temperature data for male and female is approximately normally distributed.





Now we can perform the hypothesis test on the sample data to compare the mean difference.

We take null hypothesis H_0 = Difference between mean =0

And Alternate hypothesis H_1= Difference between mean not equal to 0

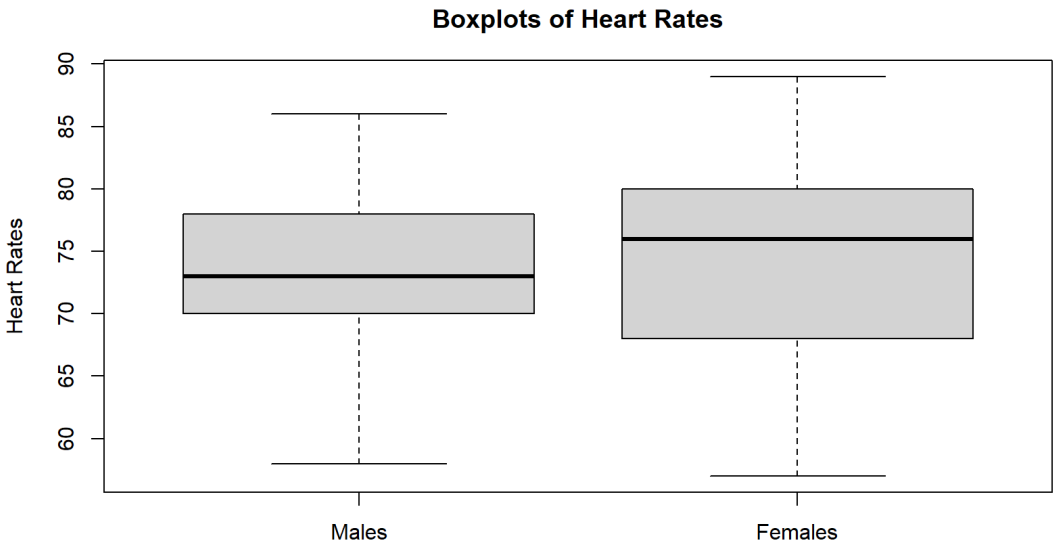
Two samples here are treated as independent samples with unequal variance and approximately normally distributed. Hence, we can use t distribution with Satterthwaite approximation to find the confidence interval.

```
data:  males$body_temperature and females$body_temperature
t = -2.2854, df = 127.51, p-value = 0.02394
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.53964856 -0.03881298
sample estimates:
mean of x mean of y
 98.10462  98.39385
```

From the test result we can observe that the CI is (-0.53964856 -0.03881298) and the p value we got is 0.02394. Since the p value does not lie in the confidence interval, we reject the null hypothesis. Hence, mean body temperature of male and female is not equal. As width of confidence interval is very small the mean difference is also very small. From the exploratory analysis we can say that mean of female body temperature is slightly higher than its counterpart.

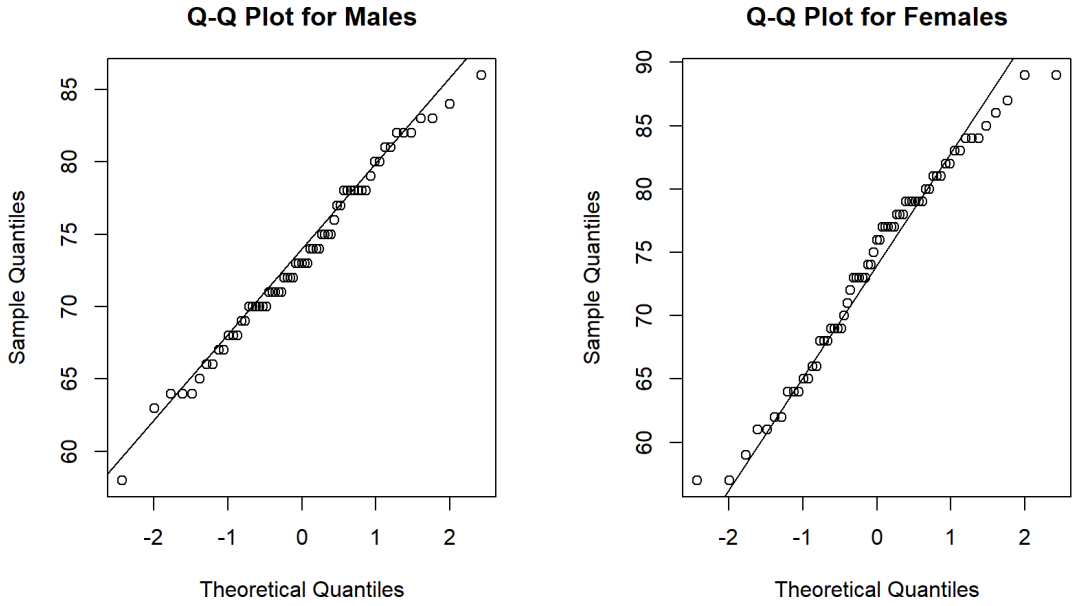
b) Do males and females differ in mean heart rate? Answer this question by performing an appropriate analysis of the data, including an exploratory analysis.

To compare mean heart rate of male and female we can plot a box plot.



From the boxplot we can observe that heart rate of female has slightly mean value and Q3 but has a lower Q1 than male. There are more outliers in female data which implies there is more variability.

Let's plot a QQ plot to see the distribution of both male and female heart rate.



We can observe that the distribution of heart rate data for male and female is approximately normally distributed.

We take null hypothesis H_0 = Difference between mean =0

And Alternate hypothesis H_1= Difference between mean not equal to 0

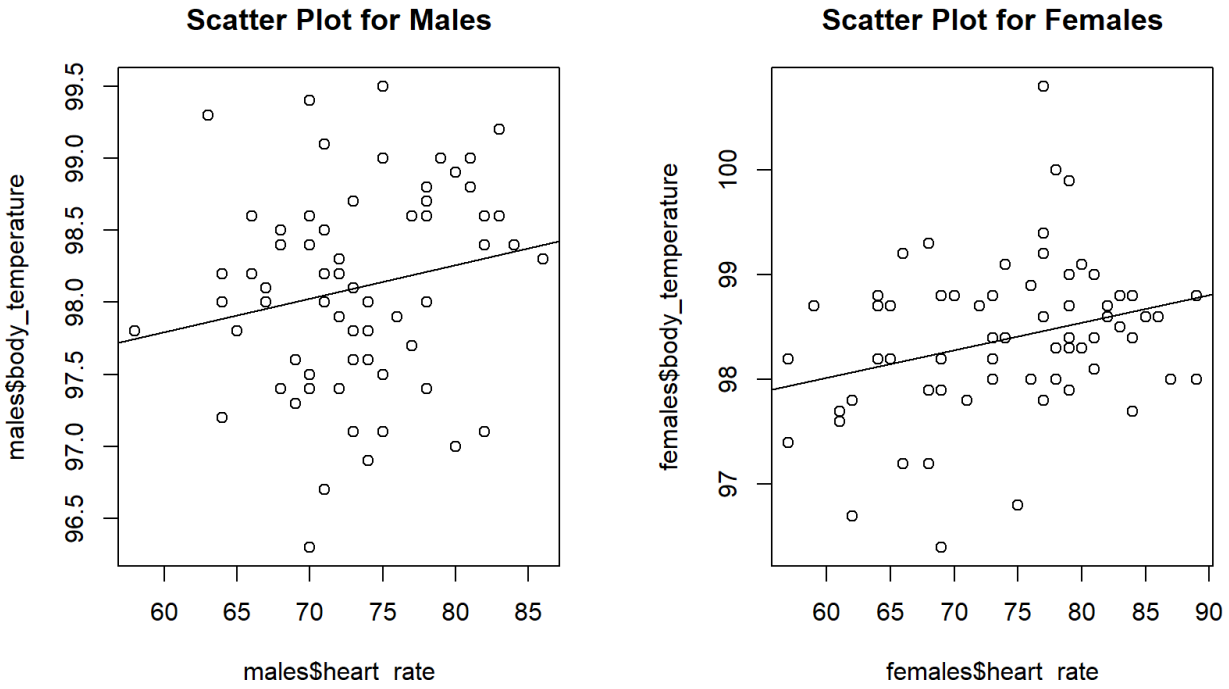
Two samples here are treated as independent samples with unequal variance and approximately normally distributed. Hence, we can use t distribution with Satterthwaite approximation to find the confidence interval.

```
data:  males$heart_rate and females$heart_rate
t = -0.63191, df = 116.7, p-value = 0.5287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.243732  1.674501
sample estimates:
mean of x mean of y
 73.36923  74.15385
```

From the test result we can observe that the CI is (-3.243732 1.674501) and the p value we got is 0.5287. Since the p value lie in the confidence interval, we accept the null hypothesis. Hence, mean heart rate of male and female is approximately equal.

c) Is there a linear relationship between body temperature and heart rate? Does this relationship depend on gender? Answer these questions by performing an appropriate analysis of the data, including an exploratory analysis.

A scatter plot can be used to find the linear relationship between body temperature and the heart rate.
Plot different scatter plots for male and female will illustrate the dependence of gender on a relation.



As we can observe both lines have a positive slope while infares that body temperature and heart rate have a positive correlation. Based on the graph we can say that it has a weak linear relationship.

Let's find the correlation values between body temperature and heart rate for both male and female.

```
> cor (males $body_temperature, males$heart_rate)
[1] 0.1955894
> cor (females $body_temperature, females$heart_rate)
[1] 0.2869312
```

Question 2

a) For a given setting, compute Monte Carlo estimates of coverage probabilities of the two intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 5000 times.

To find the coverage probability we draw different samples by simulation and compute confidence interval for each. Then the probability of coverage can be figure out by dividing total number of times the mean value lies in a confidence interval by total number of samples.
Here we are using two different approaches to calculate confidence interval. One is Z-interval, and another is percentile bootstrap.
Z-interval is calculated as:

$$\overline{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

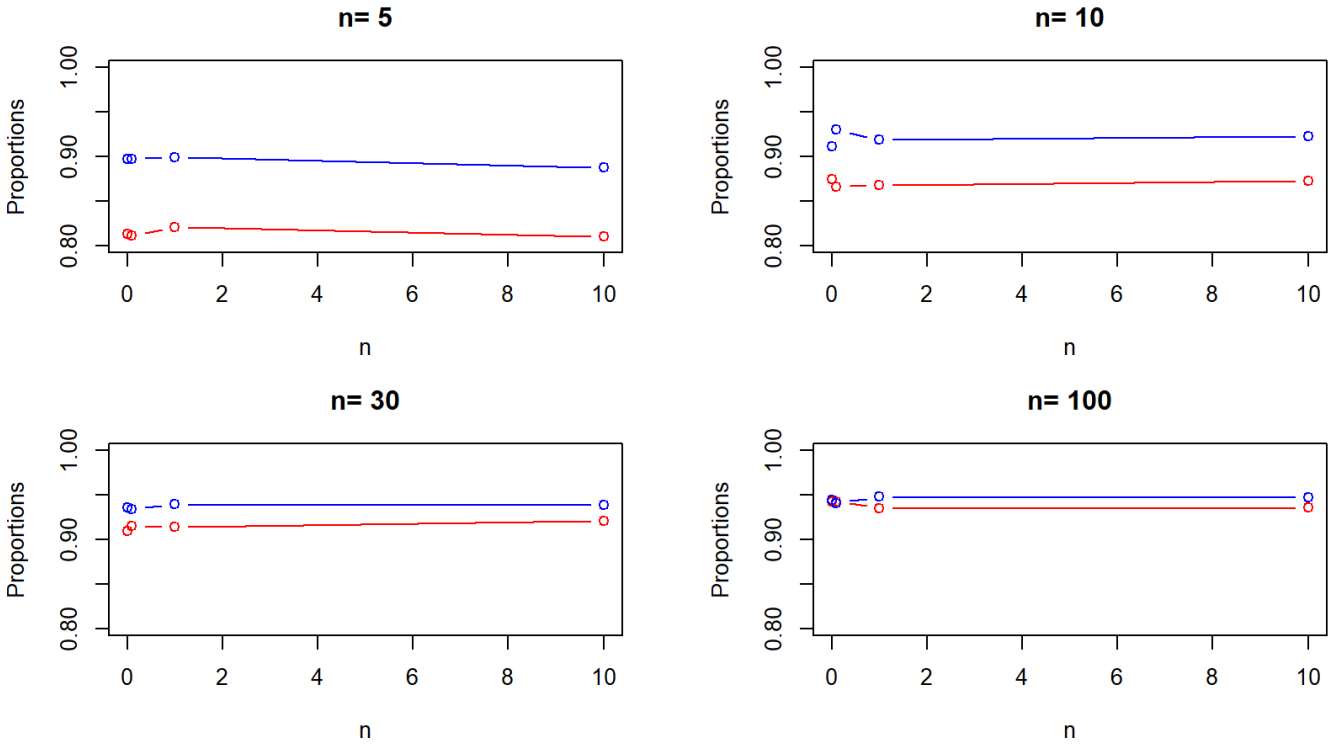
Percentile Bootstrap is calculated by taking 0.25^{th} and 0.975^{th} quartile values.

```
+ }
> #getting the value of n = 5 and lambda = 0.01 for zproportion
> zproportion(5,0.01)
[1] 0.806
> # getting the value of n = 5 and lambda = 0.01 for bproportion
> bproportion(5,0.01)
[1] 0.8892
```

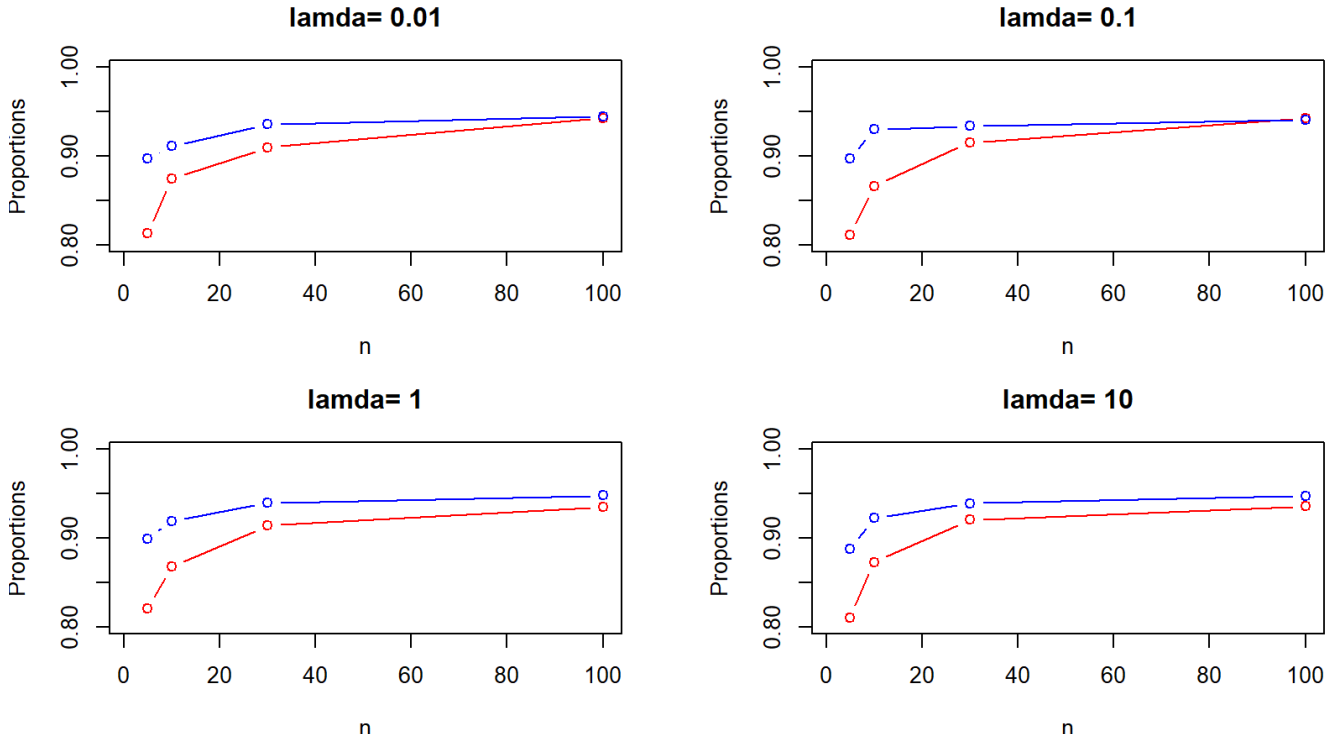
b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results.

n	lamda	zci	bci
5	0.01	0.8132	0.8974
5	0.10	0.8110	0.8974
5	1.00	0.8206	0.8994
5	10.00	0.8102	0.8878
10	0.01	0.8744	0.9110
10	0.10	0.8664	0.9298
10	1.00	0.8676	0.9188
10	10.00	0.8730	0.9228
30	0.01	0.9096	0.9358
30	0.10	0.9154	0.9340
30	1.00	0.9140	0.9394
30	10.00	0.9206	0.9390
100	0.01	0.9430	0.9448
100	0.10	0.9430	0.9410
100	1.00	0.9352	0.9480
100	10.00	0.9358	0.9474

Graphical Representation of data:



Graph 1: Red represents z-proportions and blue represents bootstrap proportions. The values are plotted against λ keeping n fixed



Graph 2: Red represents z-proportions and blue represents bootstrap proportions. The values are plotted against n keeping λ fixed

c) Interpret all the results. Be sure to answer the following questions: In case of the large-sample interval, how large n is needed for the interval to be accurate? Likewise, in case of the bootstrap interval, how large n is needed for the interval to be accurate? Do these answers depend on λ? Can we say that one method is more accurate than the other? Which interval would you recommend? Provide justification for all your conclusions.

From Graph 1, we see that the graphs don't change drastically when λ is changed, so we can say that the coverage probabilities don't depend on λ. And we also see that the coverage probabilities we get via bootstrap are higher than those of z-interval method. From Graph 2, we can conclude that the coverage probabilities depend on n. Now for the large-sample z-interval, we see that the coverage probabilities are as accurate, as the coverage probabilities we got from bootstrap method, when n is large (n=100) For the bootstrap method coverage probabilities, they are on the higher side (approximately) from n=30 onwards Taking into account all the graphs, we can say that coverage probabilities we got from bootstrap method are higher for every combination of (n,λ) than for the large-sample z-interval method, hence bootstrap method is more accurate even for the low values of n. Hence bootstrap method is recommended.

d) Do your conclusions in (c) depend on the specific values of λ that were fixed in advance? Explain.

The output from the code in Section 2 helps us to infer that the:
The coverage probability for bootstrap: n= 5 lambda= 0.1 is 0.8974
The coverage probability for large sample z: n= 5 lambda= 0.1 is 0.8132
And,
The coverage probability for bootstrap: n= 10 lambda= 0.1 is 0.9110
The coverage probability for large sample z: n= 10 lambda= 0.1 is 0.8744
The coverage probability for bootstrap: n= 30 lambda= 0.1 is 0.9358
The coverage probability for large sample z: n= 30 lambda= 0.1 is 0.9096
The coverage probability for bootstrap: n= 100 lambda= 0.1 is 0.9448
The coverage probability for large sample z: n= 100 lambda= 0.1 is 0.9430

Therefore:
The conclusions obtained in (c) hold for specific values of lambda. In this case lambda = 0.1

Section 2

Q1

```
#reading data using read.csv function
bodytemp_heartrate = read.csv("C:/Users/PRK200000/Desktop/Pranil/Download/bodytemp-heartrate.csv",header = T)
bodytemp_heartrate

#separating the two datasets
males = subset(bodytemp_heartrate, bodytemp_heartrate$gender==1)
females = subset(bodytemp_heartrate, bodytemp_heartrate$gender==2)

#drawing boxplots for body temperature values
boxplot(males$body_temperature, females$body_temperature, main = "Boxplots of Body Temperatures", names =
c('Males', 'Females'), ylab = "Temperatures")

#drawing Q-Q plots for the body temperature values
par(mfrow=c(1,2))
qqnorm(males$body_temperature, main = 'Q-Q Plot for Males')
qqline(males$body_temperature)
qqnorm(females$body_temperature, main = 'Q-Q Plot for Females')
qqline(females$body_temperature)
par(mfrow=c(1,1))

#confidence interval using t,test function for the body temperature values
t.test(males$body_temperature, females$body_temperature, alternative ='two.sided', var.equal = F)

#drawing boxplot for the heart rate values
boxplot(males$heart_rate, females$heart_rate, main = "Boxplots of Heart Rates", names= c('Males', 'Females'), ylab =
"Heart Rates")

#drawing Q-Q plot for the heart rate values
par(mfrow=c(1,2))
qqnorm(males$heart_rate, main = 'Q-Q Plot for Males')
qqline(males$heart_rate)

qqnorm(females$heart_rate, main = 'Q-Q Plot for Females')
qqline(females$heart_rate)
par(mfrow=c(1,1))

#getting the confidence interval using the t.test function
t.test(males$heart_rate, females$heart_rate, alternative = 'two.sided', var.equal = F)

#finding the correlation values between body temperatures and heart rates
cor (males $body_temperature, males$heart_rate)
cor (females $body_temperature, females$heart_rate)

#drawing the scatter plots for the body temperature and heart rate values for males and females
par(mfrow=c(1,2))
plot(males$heart_rate, males$body_temperature, pch=1, main='Scatter Plot for Males')
abline(lm(males$body_temperature~males$heart_rate))
```

```
plot(females$heart_rate, females$body_temperature, pch=1, main='Scatter Plot for Females')
abline (lm(females$body_temperature~females$heart_rate))
par(mfrow=c(1,1))
```

Q2)

```
Zinterval <- function(n, lambda) {
  x <- rexp(n,lambda)
  LowerBound <- mean(x) - qnorm(0.975) * sd(x) / sqrt(n)
  UpperBound <- mean(x) + qnorm(0.975) * sd(x) / sqrt(n)
  Mean = 1/lambda
  if(UpperBound>Mean & LowerBound<Mean) {
   return (1)
   }
  else {
   return (0)
   }
  }

#creating function zproportion
zproportion <- function(n, lambda) {
  Present_in_CI <- replicate(5000, Zinterval(n, lambda))
  Present <- Present_in_CI[which (Present_in_CI == 1)]
  return (length(Present)/5000)
  }

#getting the value of n = 5 and lambda = 0.01 for zproportion
zproportion(5,0.01)

#creating function mean.star
mean.star <- function(n,lambda) {
 u.star <- rexp(n, lambda)
 return (mean(u.star))
}
mean.star(5,0.01)
#creating function boot_CI
boot_CI <- function(n, lambda) {
  U <- rexp(n,lambda)
  TrueMean <- 1/lambda
  lambda1 = 1/mean(U)
  V <- replicate(1000, mean.star(n, lambda1))
  bound <- sort(V)[c(25, 975)]
  if(bound[2]>TrueMean & bound[1]<TrueMean) {
   return (1)
   }
  else {
   return (0)
   }
  }
#creating function bproportion
```

```
bproportion <- function(n, lambda) {
  Present_in_CI <- replicate(5000, boot_CI(n, lambda))
  Present <-Present_in_CI[which (Present_in_CI == 1)]
  return (length(Present)/5000)
  }
# getting the value of n = 5 and lambda = 0.01 for bproportion
bproportion(5,0.01)


#generating the proportion values for bootstrap and z-interval for all the combinations of n and Lamda
n = c( 5, 10, 30, 100)
lamda = c(0.01, 0.1, 1, 10)
Accuracy_matrix=matrix(nrow=0,ncol=4)
colnames(Accuracy_matrix)=c("n","lamda", "zci","bci")
for (i in n){
 for (j in lamda){
   x=c(i,j,zproportion(i,j), bproportion(i,j))
   Accuracy_matrix<-rbind(Accuracy_matrix,x)
 }
}
Accuracy_matrix


#Plotting line graph for all the combinations of n and lamda

l=0
par(mar=c(3.8,3.8,3.3,3.3))
par(mfrow=c(2,2))
for (i in n){
 plt=matrix(nrow=0,ncol=2)
 for (j in lamda){
   l=l+1
   y=c(Accuracy_matrix[l,3],Accuracy_matrix[l,4])
   plt<-rbind(plt,y)
 }
 plot(lamda, plt[,1], main = paste("n=",i,sep=" ") , xlab = 'n', ylab ='Proportions', col = 'red', type = 'b', xlim = c(0.01,10),
ylim = c(0.80,1))
 lines(lamda, plt[,2], col = 'blue', type = 'b')
}

par(mar=c(3.8,3.8,3.3,3.3))
par(mfrow=c(2,2))
for (i in 1:4){
 plt=matrix(nrow=0,ncol=2)
 for (j in seq(i,16,4)){
   y=c(Accuracy_matrix[j,3],Accuracy_matrix[j,4])
   plt<-rbind(plt,y)
 }
 plot(n, plt[,1], main = paste("lamda=",Accuracy_matrix[j-4,2],sep=" ") , xlab = 'n', ylab ='Proportions', col = 'red', type =
'b', xlim = c(1,100), ylim = c(0.80,1))
 lines(n, plt[,2], col = 'blue', type = 'b')
```

}